Analysis of Survey Data

15

Chapter Outline

All of the statistical methods we have discussed so far are based on the assumption that the data were obtained by simple random sampling with replacement. As we discussed in Chapter 6, simple random sampling can be very expensive, if not infeasible, to implement in community surveys. Consequently, survey statisticians often use alternative sample selection methods that use such design features as stratification, clustering, and unequal selection probabilities. Some of these features were briefly discussed in Chapter 6. Sample designs that use some of these more advanced design features are referred as *complex sample designs*. These complex designs require adjustments in the methods of analysis to account for the differences from simple random sampling. Once these adjustments are made, all the analytic methods discussed in this book can be used with complex survey data. We introduce several different ways of making these adjustments in this chapter, with a focus on two specific topics: the use of sample weights and the calculation of estimated variances of parameter estimates based on complex sample designs.

Our treatment of the material in this chapter differs from the treatment in the other chapters in that we provide few formulas here. Instead, we attempt to provide the reader with a feel for the different approaches. We also provide some examples pointing out how ignoring the sample design in the analysis can yield very misleading conclusions. We follow this nonformulaic path because of the mathematical complexity of the procedures. In addition, we do not go into detail about procedures for addressing two important problems in the analysis of survey data — nonresponse and missing data. There are several approaches for dealing with these problems (Levy and Lemeshow 1999; Little and Rubin 2002), but they all make assumptions about the data that are difficult to check. We cannot stress too highly the importance of reducing nonresponse in surveys. Even after reading this chapter, we think the reader will need to work with a survey statistician when carrying out the analysis of survey data.

15.1 Introduction to Design-Based Inference

There are two general approaches for dealing with the analytic complexities in survey data and these can be loosely grouped under the headings of "design-based" and "model-based." We are presenting only the design-based approach because it is the standard

way of analyzing complex surveys, although the model-based approach also has many supporters. Several sources discuss the model-based approach (Korn and Graubard 1999; Lee and Forthofer 2006; Lohr 1999).

The design-based approach requires that the sample design be taken into account in the calculation of estimates of parameters and their variances. As we just mentioned, a key feature of the complex sample design is the sample weight, which is based on the probability of selection of the units in the sample. The calculation of the estimated variance for a parameter estimate from complex survey data usually cannot be done through applying a simple formula. The following special procedures are used.

15.2 Components of Design-Based Analysis

As just mentioned, most community surveys utilize complex sample designs to facilitate the conduct of the surveys. As a result of using stratification and clustering, the selection probabilities of units are unequal. In some surveys, unequal selection probabilities are used intentionally to achieve certain survey objectives. For example, the elderly, children, and women of childbearing ages are often oversampled to obtain a sufficient number of people in those categories for detailed analysis.

15.2.1 Sample Weights

The weight is used to account for differential representation of sample observations. The weight is defined as the inverse of selection probability for each observation. Let us explore the concept of the sample weight in the simple random sampling situation. Suppose that an SRS of $n = 100$ households was selected from a population of $N = 1000$ households to estimate the total medical expenditure for a year for the population. The selection probability of each sample observation is $n/N = 0.1$, and the sample weight is therefore 10 ($= N/n$). The sample weights add up to N. If the average annual medical expenditure for the sample ($\bar{y}$) was found to be \$2000, then the estimated total medical expenditure for the population would be $N\bar{y}$ = \$2,000,000. Another way of writing the estimate is

$$N\bar{y} = N\left(\frac{\sum y_i}{n}\right) = \sum\left(\frac{N}{n}\right)y_i$$

or the weighted total of sample observations. Since the weight is the same for all sample observations in simple random sampling, we don't need to weight each observation separately.

The situation is slightly different with a disproportionate stratified random sample design. Suppose the population of 1000 households consists of two strata: 200 (N_1) households with at least one senior citizen and 800 (N_2) households without any seniors. Suppose further that 50 households were randomly selected from each stratum. The selection probability in the first stratum is 50/200 and the weight is 4 ($= N_1/n_1$). In the second stratum the selection probability is 50/800 and the weight is 16. If the average medical expenditure for the first and second stratum were found to be \$5000 ($\bar{y}_1$) and \$1250 ($\bar{y}_2$), respectively, then the estimated total medical expenditure for the population would be \$2,000,000 ($= N_1\bar{y}_1 + N_2\bar{y}_2 = 200\{1250\} + 800\{5000\}$). The following relationship shows the role of the weight in estimation:

$$N_1\bar{y}_1 + N_2\bar{y}_2 = N_1\left(\frac{\sum y_{1i}}{n_1}\right) + N_2\left(\frac{\sum y_{2i}}{n_2}\right) = \sum\left(\frac{N_1}{n_1}\right)y_{1i} + \sum\left(\frac{N_2}{n_2}\right)y_{2i}.$$

Although we have used SRS and stratified sample designs to introduce the sample weights, the same concept extends to more complex designs. When each observation in the sample has a different weight (w_i), the estimates for the population can be obtained using the following general estimator:

$$\text{population estimate} = \sum w_i y_i.$$

This procedure applies to all sample designs.

In survey analysis, the weight is often modified further by poststratification adjustments discussed in the following sections.

15.2.2 Poststratification

In the health field, many of the variables of interest vary by, for example, a person's age, sex, and race. If we knew these variables before we carried out the survey, we could use them to create a stratified design that would take these variables into account. Unfortunately, we often don't know the values of these variables before the survey, and this fact prevents us from using a stratified sample design.

However, we still wish to take these variables into account in the analysis. We do this by adjusting the sample distributions so that they match their population distributions for these variables. We accomplish this matching by using a technique called poststratification that adjusts the sample weights after (post) the sample data have been collected.

The following example shows how poststratification adjustment is created for the different categories.

Example 15.1

A telephone survey was conducted in a community to estimate the average amount spent on food per household in a week. Telephone surveys are popular because they are quick and easy to perform. Unfortunately, they exclude the small percentage of the households without a landline telephone, and this exclusion could introduce some small degree of bias in the results. With the increasing use of cell phones, the potential for bias in telephone surveys is increasing unless cell phone numbers are also included. Given the goal of this survey, one desirable stratification variable would be household size because larger households likely spend more than smaller households. Since information on household size was not readily available before the survey was conducted, we could not stratify on this variable in the survey design.

The survey failed to obtain responses from 12 percent of the households. It was thought that these nonrespondents were more likely to be living in smaller households, and this idea is supported by the data shown in Table 15.1. These data show the distribution of sample households by household size and the corresponding distribution from a previous study involving household size in the community. Smaller

Table 15.1 Poststratification adjustments by household size for a telephone survey.

Number of Persons in Household	Number of Households in Sample	Sample Distribution	Population Distribution	Adjustment Weight[a]	Average Food Expenditure
1	63	0.2072	0.2358	1.13803	$38
2	81	0.2664	0.3234	1.21396	52
3	63	0.2072	0.1700	0.82046	78
4	52	0.1711	0.1608	0.93980	98
5+	45	0.1480	0.1100	0.74324	111

[a]Population distribution divided by sample distribution

households are indeed underrepresented in the sample and this suggests the average food expenditure would be overestimated unless we make an adjustment for household size. Table 15.1 shows the procedure of poststratification adjustment.

The postratification adjustment for single-person households is to multiply the number of single-person households by 1.138, reflecting that this category of households is underrepresented by 14 percent. This adjustment is equivalent to multiplying the sample weights by the same factor. The adjusted number of households for this category is then 71.7 (= 63{1.138}); for the rest of household size categories the adjusted numbers are 98.3, 51.7, 48.9, and 33.4. The distribution of these adjusted numbers of household by household size now matches the distribution in the population. The average food expenditure based on these adjusted numbers of households is $67.00, compared with the unadjusted average of $71.08. As expected, the adjusted average is lower than the unadjusted estimate.

We have not addressed the nonresponders directly through the poststratification adjustment. Given that the nonresponse rate was only 12 percent, it is unlikely that the average food expenditure estimate would change much if the nonresponders were included. However, it would be good to do more follow-up with a sample of the nonresponders in an effort to determine if they differed drastically from the responders.

Multiplying the sample weight by the poststratification adjustment factors causes the weighted sample distribution to match the population distribution for the variables used in the poststratification.

15.2.3 The Design Effect

In Chapter 6 we demonstrated in one example that a stratified sample could provide a more precise — have smaller sampling variance — estimator for the sample mean than a simple random sample of the same sample size. In this section we provide a measure, the *design effect*, for comparing a sample design to a simple random sample design with replacement. To introduce this idea, we will begin by comparing simple random sampling without replacement to simple random sampling with replacement.

In Chapters 7 and 8, we used s^2/n as the estimator for the variance of the sample mean ($\bar{x}$) and $s/\sqrt{n}$ as the estimator for the standard error for data resulting from

simple random sampling with replacement. When simple random sampling without replacement is used, the formula for the estimated variance is

$$\hat{\mathrm{V}}\mathrm{ar}(\bar{x}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right). \tag{15.1}$$

The term $(1 - n/N)$, called the *finite population correction* (FPC), adjusts the formula to take into account that we are no longer sampling from an infinite population. Use of this term decreases the magnitude of the variance estimate. For samples from large populations, the FPC is approximately one, and it can be ignored in these cases.

The ratio of the sampling variance of SRSWOR to that of SRSWR is the FPC, and it reflects the effect of using SRSWOR compared to using SRSWR. This ratio comparing the variance of some statistic from any particular sample design to that of SRSWR is called the *design effect* for that statistic. It is used to assess the loss or gain in precision of sample estimates from the sample design used. A design effect less than one indicates that fewer observations are needed to achieve the same precision as SRSWR whereas a design effect greater than one implies that more observations may be needed to yield the same precision. Extending this concept to sample size, the *effective sample size* of a design is the size of a simple random sample with replacement that would have produced the same estimated sample variance for the estimator under consideration. The effective sample size is the actual sample size of the design being used divided by the design effect.

The design effect can be examined theoretically for some simple sample designs. As was just mentioned, we pointed out in Chapter 6 that stratified random sampling often produces smaller sampling variance than SRS. Cluster sampling will lead to a greater sampling variability when the sampling units are similar within clusters. The *intraclass correlation coefficient* (ICC) is used to assess the variability within the clusters. The ICC is the Pearson correlation coefficient based on all possible pairs of observations within a cluster.

The design effect of single-stage cluster sample design with equal size clusters is

$$1 + (M - 1)ICC$$

where M is the size of each cluster. Given this design, the ICC ranges from $-1/(M - 1)$ to 1. When ICC is positive, the design effect will be greater than one. If the clusters were formed at random, then ICC = 0; when all the units within each cluster have the same value, then ICC = 1 and the design effect is the same as the size of the cluster. Most clusters used in community surveys consist of houses in the same area, and these generally yield positive ICCs for many survey variables such as socioeconomic and some demographic characteristics.

Since the determination of the design effect requires that we have an estimate of the sample variance for a given design, this calculation is usually not a simple task for a complex sample design. The complexity of the design often means that we cannot use the variance estimating formulas presented in previous chapters; rather, special techniques that utilize unfamiliar strategies are required. The next section presents several strategies for estimating sampling variance for statistics from complex sample designs.

15.3 Strategies for Variance Estimation

The estimation of the sampling variance of a survey statistic is complicated not only by the complexity of the sample design but also by the form of the statistic. Even with an SRS design, the variance for some sample statistics requires nonstandard estimating techniques. For example, the sampling variance of sample median was not covered in previous chapters. Moreover, the variance estimator for a weighted statistic is complicated because both the numerator and denominator are random variables. We will present several techniques for estimating sampling variances: (1) from complex samples and (2) for nonlinear statistics. These techniques include replicated sampling, balanced repeated replication, jackknife repeated replication, and the linearization method (Taylor series approximation).

15.3.1 Replicated Sampling: A General Method

The replicated sampling method requires the selection of a set of subsamples from the population with each subsample being drawn independently following the same sample selection design. Then an estimate is calculated for each subsample, and the sampling variance of the overall estimate based on all the subsamples can be estimated from the variability of these independent subsample estimates. The repeated systematic sampling discussed in Chapter 6 represents this strategy.

The standard error of the mean ($\bar{u}$) of t replicate estimates, $u_1, u_2, \ldots, u_1$ of the parameter U can be estimated by

$$\sqrt{\sum_{i=1}^{t}(u_i-\bar{u})^2/[t(t-1)]}. \tag{15.2}$$

This estimator can be applied to any sample statistic obtained from independent replicates for any sample design.

In applying this estimator, ten replicates are recommended by Deming (1960) and a minimum of four by Sudman (1976) for descriptive statistics. An approximate estimate of the standard error can be calculated by dividing the range in the replicate estimates by the number of replicates when the number of replicates is between 3 and 13 (Kish 1965). However, because this estimator with t replicates is based on $t - 1$ degrees of freedom, a larger number of replicates may be needed for analytic studies, perhaps 20 to 30 (Kalton 1983).

Example 15.2

In this artificial example, we demonstrate the use of replicated sampling for the estimation of the sample variance of a statistic. In this case, we are going to estimate the population proportion of male births and the sample variance of this statistic based on replicated samples. Instead of collecting actual data, we will use the random digits in Table B1 to create our replicated samples. In our simulation process, we are going to assume that the population proportion of male births is 0.5. We will take 10 replicated samples of size 40 using the first eight 5-digit-columns of lines 1

Table 15.2 Estimation of standard errors for the proportion of boys from 10 replicated samples of size 40.

Replicate	*n*	Number of Boys	Proportion of Boys		Standard Error
Full sample	400	205	0.512		0.025[a]
1	40	21	0.525		
2	40	16	*0.400*		
3	40	21	0.525		
4	40	20	0.500		
5	40	20	0.500		
6	40	17	0.425		
7	40	21	0.525		
8	40	26	*0.650*		
9	40	24	0.600		
10	40	19	0.475		0.022[b]
			(0.650 – 0.400)/10	=	0.025[c]

[a]Based on SRS
[b]Based on Equation (15.2)
[c]Based on the range in replicate proportions divided by the number of replicates

through 10. Table 15.2 shows the number and proportion of boys with estimates of the standard error by three different methods.

For the full sample of 400 — combining the data from the 10 separate samples — the proportion of boys is 0.512 and its standard error is 0.025 based on simple random sampling. The standard error estimated from the 10 replicate estimates using Equation (15.2) is 0.022. An approximate estimate can also be obtained by taking the range in replicate estimates divided by the number of replicates. This value is 0.025 ([0.650 – 0.400]/10). Of course replicated sampling is not needed for estimating standard errors for a simple random sample design. But this strategy also works for complex sample designs.

The chief advantage of replicated sampling is the ease in estimation of the standard errors for complex sample designs. This strategy is especially useful in systematic sampling, since there is no way to estimate the standard error of an estimator from a systematic sample with only one replicate. Replicated systematic sampling can easily be implemented by randomly selecting multiple starting points. In applying Equation (15.2), the sample statistic for the full sample is generally used instead of the mean of replicate estimates when sample weights are present.

However, replicated sampling is difficult to implement in multistage cluster sampling designs and is seldom used in large-scale surveys. Instead, the replicated sampling idea can be applied in the data analysis stage where pseudo-replication methods for variance estimation are used. The next two sections present two such methods.

15.3.2 Balanced Repeated Replication

The balanced repeated replication (BRR) method represents an application of the replicated sample idea to a paired selection design in which two primary sampling units (PSU) are sampled from each stratum. The paired selection design is often used to simplify the calculation of variance within a large number of strata. The variance

between two units within a stratum is one-half of the squared difference between the units. McCarthy (1966) originally proposed the BRR method for the National Center for Health Statistics for analyzing the National Health Examination Survey that used a paired selection design.

To apply the replicated sampling idea, half-sample replicates are created by taking one PSU from each stratum. From the paired selection design, we can create only two half-sample replicates. Since the estimate of standard error based on two replicates is unstable, we repeat the process of forming half-sample replicates in such a way that replicates are independent of each other (Plackett and Burman 1946).

Replicate estimates, $u_1, u_2, \ldots, u_t$, for a sample statistic are calculated by doubling the sample weights, since each replicate contains one-half of the total observations. Then the standard error of the statistic ($\bar{u}$) for the full sample can be calculated by

$$\sum_{i=1}^{t} (u_i - \bar{u})^2 / t.$$

Since this process involves so much manipulation of the data, it is usually necessary to use specialized computer software created to carry out the BRR approach.

15.3.3 Jackknife Repeated Replication

Another replication-based procedure is called the jackknife repeated replication method (JRR). This procedure creates pseudo-replicates by deleting one unit from the sample, then calculating the sample statistic of interest. That unit is put back into the sample, another unit is deleted and the statistic is calculated, and so on. The estimate of the variance is then based on the variation in these sample statistics. The term *jackknife* may be used because this procedure can be used for a variety of purposes. The idea of jackknifing was introduced by Quenouille in 1949 in the estimation of bias for a sample estimator. Frankel (1971) first applied the jackknife procedure to the computation of sampling variance in complex surveys, using it in a manner similar to the BRR method. The following example illustrates the principle of jackknifing.

Example 15.3

We consider a small data set — the ages of the first 10 patients in DIG40 shown in Table 3.1. Assuming that these data are from a simple random sample, the sample mean is 58.2 and the sample median is 59.5. If we ignore the FPC, the estimated standard error of the sample mean is 4.27. These statistics are shown in Table 15.3 along with the 10 observations. We next estimate the standard error of the sample mean by the jackknife procedure.

We create a jackknife replicate by deleting the first observation (age 55) and calculate the mean for the replicate, which gives 58.56, as shown in the table. By deleting the second observation (age 78) we get the second jackknife replicate estimate of 56. Repeating the same procedure we have 10 replicate estimates, $\bar{y}_{(1)}, \bar{y}_{(2)}, \ldots, \bar{y}_{(10)}$. Let the mean of the replicate estimates be $\bar{\bar{y}} (= \sum \bar{y}_{(i)}/n)$, and this value is 58.2,

Table 15.3 Estimation of standard error by the jackknife procedure for the mean and median age for 10 patients in DIG40.

		Jackknife Replicate Estimates	
Patient	**Age**	**Mean**	**Median**
1	55	58.56	60
2	78	56.00	59
3	50	59.11	60
4	60	58.00	59
5	31	61.22	60
6	70	56.89	59
7	46	59.56	65
8	59	58.11	59
9	68	57.11	59
10	65	57.44	59
Mean	58.2	58.2	59.9
Median	59.5		
Standard error estimates for the mean:			
From the sample	4.27		
From jackknife replicates	4.27		
Standard error estimate for the median:			
From jackknife replicates	5.27		

which is the same as the sample mean. The standard error can be estimated by $\sqrt{(n-1)\Sigma(\overline{y}_{(i)}-\overline{\overline{y}})^2/n}$, which equals 4.27. The standard error estimated from replicate estimates is the same as the estimate obtained directly from the sample, suggesting the jackknife procedure works.

The jackknife procedure also allows us to estimate the standard error for the median. The first replicate estimate for the median is based on the nine observations remaining after deleting the first observation as before. Deleting each observation in turn allows us to determine ten replicate estimates of the median as shown in Table 15.3. The mean of the replicate medians is 59.9. Using the same formula shown above, we can estimate the standard error of the median and this value is 5.27.

For a complex sample design, the JRR method is generally applied at the PSU level. The JRR method is not restricted to a paired selection design but is applicable to any number of PSUs per stratum. Let us consider a situation with L strata. If u_{hi} is the estimate of the parameter U from the hth stratum and ith replicate, n_h is the number of sampled PSUs in the hth stratum, and r_h be the number of replicates formed in the hth stratum, then the standard error is estimated by

$$\sqrt{\sum_{h=1}^{L}\left(\frac{n_h-1}{r_h}\right)\sum_{i=1}^{r_h}(u_{hi}-\overline{u})^2}.$$

If each of the PSUs in the hth stratum is removed to form a replicate, r_h is the same as n_h in each stratum, but the formation of n_h replicates in the hth stratum is not required. When the number of strata is large and n_h is two or more, we can reduce the computational burden by using only one replicate in each stratum. However, a sufficient number

of replicates must be used in analytic studies to ensure that there are adequate degrees of freedom.

15.3.4 Linearization Method

A completely different approach from the pseudo-replication methods for estimating variances from complex survey designs follows a more mathematical approach. This mathematical approach, called Taylor series linearization, is used in statistics to obtain a linear approximation to nonlinear functions. The beauty of the Taylor series is that many nonlinear functions are approximated quite well by only the first few terms of the series. This approach has gained wide acceptance in the analysis of weighted data from complex surveys because many of the statistics that we estimate, including regression coefficients, are nonlinear, and their estimated variances are also nonlinear. This approach to variance estimation has several other names in the literature, including the linearization method and the delta method. A brief presentation of the Taylor series approach and an example is presented in Appendix A.

The following example demonstrates how the linearization works for the calculation of sampling variance of a *ratio estimate*.

Example 15.4

We consider a small sample for this illustration. A simple random sample of eight health departments was selected from 60 (N) rural counties to estimate the total number of professional workers with a master of public health degree. It is known that the 60 health departments employ a total of 1150 (Y) professional workers. The sample data shown in Table 15.4 are the number of professional workers (y_i) and the number of professional workers with an MPH degree (x_i).

Based on the sample data on the number of workers with an MPH degree, we can estimate the total number of professional workers with an MPH degree, that is 630 ($= N\bar{x} = 60 * 10.5$). The variance of this estimate is $\hat{Var}(N\bar{x}) = N^2\hat{Var}(\bar{x})$ as shown in Chapter 4. Using Equation 15.1 for $\hat{Var}(\bar{x})$, the estimated standard error of this estimate is

$$\sqrt{\frac{N^2\sum(x_i-\bar{x})^2}{n(n-1)}\left(1-\frac{n}{N}\right)} = 97.3.$$

Table 15.4 Numbers of professional workers and professionals with an MPH degree for 8 sample health departments.

Health Department	Number of Professional Workers (y_i)	Number of Workers with MPH (x_i)
1	21	14
2	18	8
3	9	3
4	13	6
5	15	8
6	22	13
7	30	17
8	27	15
Mean	19.375	10.5

Since the total number of professional workers for the population is known and x and y are highly correlated, we prefer to use a ratio estimate. The ratio estimate of the total number of professional workers with an MPH is

$$\hat{X} = \left(\frac{\bar{x}}{\bar{y}}\right) Y = \left(\frac{10.5}{19.375}\right) 1150 = 623.$$

The standard error calculation for the ratio estimate is complicated because both the numerator and denominator of the ratio are random variables. The linearization method can provide an approximation. Using Stata we obtained the following results (see **Program Note 15.1** on the website):

Ratio	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
totxmph/prof	623.2258	29.92157	552.4725	693.9791	1

The estimated standard error for the ratio estimate is 29.9, which is much smaller than that obtained by simple random sample estimator (97.3), suggesting that the ratio estimate is the preferred method of estimation for this case.

The Taylor series approximation is applied to PSU totals within strata — that is, the variance estimate is a weighted combination of the variation across PSUs within the same stratum. This calculation is complex but can require much less computing time than the replication methods just discussed. This method can be applied to any statistic that is expressed mathematically — for example, the mean and the regression coefficient. But it cannot be used with nonfunctional statistics such as the median and other percentiles.

15.4 Strategies for Analysis

We introduce the Third National Health and Nutrition Examination Survey (NHANES III) here to illustrate the methods of survey data analysis. NHANES III, sponsored by NCHS (1994), collected information on a variety of health-related subjects from a large number of individuals through personal interviews and medical examinations. Its sample design was complex to accommodate the practical constraints of cost and survey requirements, resulting in a stratified, multistage, probability cluster sample of eligible persons in households. The PSUs were counties or small groups of contiguous counties and a total of 2812 PSUs were divided into 47 strata based on demographic characteristics. Thirteen of the 47 strata contained one large urban county, and these urban PSUs were automatically included in the sample. Two PSUs were sampled from each of the remaining 34 strata. The subsequent hierarchical sampling units included census enumeration districts, clusters of households, households, and eligible persons. Preschool children, the aged, and the poor were oversampled to provide sufficient numbers of persons in these subgroups. The NHANES III was conducted in two phases. The 13 large urban counties were rearranged into 21 survey sites, subdividing some large counties. Combining with nonurban PSUs, 89 survey sites were randomly divided into two sets: 44 sites were surveyed in 1988–1991 (Phase I) and the remaining 45 sites in 1991–1994 (Phase

II). Each phase sample can be considered an independent sample, and the combined sample can be used for a large-scale analysis.

We chose to use the Phase II adult sample (17 years of age and over) of NHANES III. It included 9920 observations that are arranged in 23 pseudo-strata with 2 pseudo-PSUs in each stratum. The sample weight contained in the public-use micro data files is the expansion weight (inverse of selection probability adjusted for nonresponse and poststratification). We created a working data file by selecting variables and calculating new variables such as body mass index. The expansion weight was converted to the relative weight by dividing the expansion weight by the average weight.

15.4.1 Preliminary Analysis

Survey data analysis begins with a preliminary exploration to see whether the data are suitable for a meaningful analysis. One important consideration in the preliminary examination of sample survey data is to examine whether there is a sufficient number of observations available in the various subgroups to support the proposed analysis. Based on the unweighted tabulations, the analyst determines whether sample sizes are large enough and whether categories of the variables need to be collapsed. The unweighted tabulations also give the number of the observations with missing values and those with extreme values, which could indicate either measurement errors or errors of transcription.

It is also necessary to examine if all the PSUs have a sufficient number of observations to support the planned analysis. Some PSUs may contain only a few or no observations because of nonresponse and exclusion of missing values. If necessary, the PSUs with none or only a few observations may be combined with an adjacent PSU within the same stratum. If a stratum contains only a single PSU as a result of combining PSUs, it may be combined with an adjacent stratum. However, collapsing too many PSUs and strata is not recommended because the resultant design may now differ substantially from the original design.

The number of observations that is needed in each PSU is dependent on the type of analysis planned. The required number is larger for analytic studies than for estimation of descriptive statistics. A general guideline is that the number should be large enough to estimate the intra-PSU variance for a given estimate.

The first step in a preliminary analysis is to explore the distributions of key variables. The tabulations may point out the need for refining operational definitions of variables and for combining categories of certain variables. Based on summary statistics, one may discern interesting patterns in the distributions of certain variables in the sample. After analyzing the variables one at a time, we can use standard graphs and SRS-based statistical methods to examine relations among variables. However given the importance of sampling weights in survey data, any preliminary analysis ignoring the weights may fail to uncover important aspects of the data.

One way to conduct a preliminary analysis taking weights into account is to select a subsample of manageable size with the probability of selection proportional to the magnitude of the weights (PPS). The PPS subsample can be explored with the regular

descriptive and graphic methods, since the weights are now reflected in the selection of the subsample.

Example 15.5

For a preliminary analysis, we generated a PPS sample of 1000 from the 9920 persons in the adult file of Phase II of NHANES III. We first sorted the total sample by stratum and PSU and then selected a PPS subsample systematically using a skipping interval of 9.92 on the scale of cumulative relative weights. The sorting by stratum and PSU preserved in essence the integrity of the original sample design.

Table 15.5 demonstrates the use of our PPS subsample analyzed by conventional statistical methods. In this demonstration, we selected several variables that are likely to be most affected by the weights. Because of oversampling of the elderly and ethnic minorities, the weighted estimates are different from the unweighted estimates for mean age and percent Hispanic. The weights also make a difference for vitamin use and systolic blood pressure as well as for the correlation between body mass index and systolic blood pressure. The subsample estimates, although not weighted, are very close to the weighted estimates in the total sample, supporting the use of a PPS subsample for preliminary analysis.

Table 15.5 Comparison of sample statistics based on the PPS subsample and the total sample, NHANES III, Phase II (adults 17 years of age and older).

	Sample Statistics				
Sample	Mean Age	Percent Hispanic	Mean SBP[a]	Percent Vitamin Use	Correlation BMI[b] & SBP
PPS subsample (n = *1000)*					
Unweighted	42.9	5.9	122.2	43.0	0.235
*Total sample (*n = *9920)*					
Weighted	43.6	5.4	122.3	42.9	0.243
Unweighted	46.9	26.1	125.9	38.4	0.153

[a]Systolic blood pressure
[b]Body mass index

15.4.2 Subpopulation Analysis

When we analyze the data from a simple random sampling design, it is customary to perform some specific subdomain analysis — that is, to analyze separately, for example, different age groups or different sexes. However we have to be careful how we carry out this practice with complex survey data. Elimination of observations outside the specific group of interest — say, Hispanics, for example — does not alter the sample weights for Hispanics, but it can complicate the calculation of variances. For example, selecting Hispanics for analysis may mean that there are a small number or even no observations in some PSUs. As a result, several PSUs and, possibly, even strata might have to be combined to be able to calculate the variances. However, the sample structure resulting from these combinations may no longer resemble the original sample design. Thus, selecting out observations from a complex survey sample may lead to an incorrect estimation of variance (Korn and Graubard 1999, Section 5.4). The correct estimation

of variance requires keeping the entire data set in the analysis and assigning weights of zero to observations outside the group of interest.

Example 15.6

Let us consider the case of estimating the mean BMI for African Americans from Phase II of NHANES III. For illustration purposes, we attempted to select only African Americans from the sample, but we could not carry out the analysis because the computer program we were using detected PSUs with no observations. A tabulation of African Americans by stratum and PSU showed that only one PSU remained in the 13th and 15th strata. After collapsing these two strata with adjacent strata (arbitrarily with the 14th and 16th stratum, respectively), we obtained the mean BMI of 27.25 with the design effect of 2.78.

The subpopulation analysis using the entire sample and assigning weights of zero to non–African American observations produced the same sample mean BMI of 27.25, but the design effect was now 1.07, a much smaller value. For the use of subpopulation analysis, see **Program Note 15.2** on the website.

15.5 Some Analytic Examples

This section presents various examples based on Phase II of NHANES III data. The emphasis is on the demonstration of the effects of incorporating the sample weights and the design features on the analysis, rather than examining substantive research questions. We begin with descriptive analysis followed by contingency table analysis and regression analysis.

15.5.1 Descriptive Analysis

In descriptive analysis of survey data, the sample weights are used, and the standard errors for the estimates are calculated using one of the methods discussed that incorporate strata and PSUs. When the sample size is small, the FPC is also incorporated in the calculation of the standard errors. The method of calculating confidence intervals follows the same principles shown in Chapter 7. However, the degrees of freedom in the complex sample design are the number of PSUs sampled minus the number of strata used instead of $n - 1$. In certain circumstances, the determination of the degrees of freedom differs from this general rule (Korn and Graubard 1999, Section 5.2).

Example 15.7

We calculated sample means and proportions for selected variables from Phase II of NHANES III. We incorporated the sample weights, strata, and PSUs in the analysis, but the FPC was not necessary because the sample size was 9920. Table 15.6 shows the weighted and unweighted estimates and the standard errors, 95 percent confidence intervals, and the design effects for the weighted estimates.

Table 15.6 Descriptive statistics for selected variables: adult sample, Phase II of NHANES III (n = 9920).

Variable	Unweighted Statistics	Weighted Statistics	Standard Error	Confidence Interval	Design Effect
Mean age (years)	46.9	43.6	0.57	(42.4, 44.7)	10.31
Percent Black	29.8	11.2	0.97	(9.2, 13.3)	9.42
Percent Hispanic	26.1	5.4	0.71	(4.0, 6.9)	9.68
Mean years of education*	10.9	12.3	0.12	(12.1, 12.6)	15.01
Mean SBP (mmHg)*	125.9	122.3	0.39	(121.4, 123.0)	4.20
Mean BMI*	26.4	25.9	0.12	(25.7, 26.2)	5.00
Percent vitamin use*	38.4	43.0	1.22	(40.4, 45.5)	5.98
Percent smoker*	46.2	51.1	1.16	(48.7, 53.5)	5.28

*A small number of missing values were imputed.

The differences between the weighted and unweighted estimates are large for several variables. The weighted mean age is about 3.5 years smaller than the unweighted mean reflecting the oversampling of the elderly. The weighted proportion of blacks is over 60 percent smaller than the unweighted proportion and the weighted proportion of Hispanics is nearly 80 percent smaller than the unweighted, reflecting the oversampling of these two ethnic groups. The weighted mean years of education is nearly two years greater than the unweighted mean, reflecting that the oversampled elderly and/or minority groups have lower years of schooling. The weighted percent of vitamin use is also somewhat greater than the unweighted estimate.

The standard errors for the weighted estimates were calculated by the linearization method. The design effects shown in the last column suggest that the estimated standard errors are considerably greater than those calculated under the assumption of simple random sampling. The 95 percent confidence intervals for the weighted estimates were calculated using the t value of 2.0687 based on 23 (= 46 PSUs − 23 strata) degrees of freedom. See **Program Note 15.3** for this descriptive analysis.

15.5.2 Contingency Table Analysis

In Chapter 10, we used the Pearson chi-square statistic to test the null hypothesis of independence in a contingency table under the assumption that data came from an SRS. For the analysis of a two-way table based on complex survey data, the test procedure needs to be changed to account for the survey design. Several different test statistics have been proposed. Koch et al. (1975) proposed the use of the Wald statistic and it has been used widely. The Wald statistic is usually converted to an F statistic to determine the p-value. In the F statistic, the numerator degrees of freedom are tied to the dimension of the table and the denominator degrees of freedom reflect the survey design.

We illustrate the use of Wald statistic based on a 2 by 2 table examining the gender difference in prevalence of asthma based on data from Phase II of NHANES III. We first look at the unweighted tabulation of asthma by sex shown in Table 15.7. Ignoring the sample design, the prevalence rates for males and females are 6.1 and 7.6 percent, respectively. The Pearson chi-square value and the associated p-value shown in the table mean that the difference between the two prevalence rates is statistically significant at

Table 15.7 Unweighted tabulation of asthma by sex: Phase II, NAHNES III.

Asthma	Male	Female	Total
Present (Percent)	264 (6.1)	421 (7.6)	685 (6.9)
Absent	4085	5150	9235
Total	4349	5571	9920
	Chi-square (1):	8.397	
	p-value:	0.004	

Table 15.8 Weighted proportions for asthma by sex, Phase II, NAHNES III.

Asthma	Male	Female	Total
Present	0.0341 (p_{11})	0.0445 (p_{12})	0.0786 ($p_{1\cdot}$)
Absent	0.4440 (p_{21})	0.4775 (p_{22})	0.9214 ($p_{2\cdot}$)
Total	0.4781 ($p_{\cdot 1}$)	0.5219 ($p_{\cdot 2}$)	1.0000 ($p_{\cdot\cdot}$)
	Wald statistics:	Chi-square: 3.2941	
		F (1, 23): 3.2941	
		p-value: 0.0826	

the 0.01 level. However, we know this conclusion could be misleading because we did not account for the sample design in the calculation of the test statistic.

Let us now look at weighted cell proportions shown in Table 15.8. Under the null hypothesis of independence, the estimated expected proportion in cell (1, 1) is $(p_1)(p_1)$. Let $\hat{\theta} = p_{11} - (p_1)(p_1)$. Then Wald chi-square is defined as

$$X_w^2 = \hat{\theta}^2 / \hat{V}(\hat{\theta}).$$

We can find $\hat{V}(\hat{\theta})$ by using one of the methods discussed in previous section. The Wald test statistic, X_w^2, approximately follows a chi-square distribution with one degree of freedom.

For the weighted proportions in Table 15.8, $\hat{\theta} = -0.0034786$ and its variance is 0.000003674 (calculated using the linearization method). The Wald chi-square is

$$\frac{\hat{\theta}^2}{\hat{V}\left(\hat{\theta}\right)} = \frac{(-0.0034786)^2}{0.000003674} = 3.2941$$

and the associated p-value is 0.070. A more accurate p-value can be obtained from $F(1, 23) = 3.2941$ with p-value of 0.083. Taking into account the sample design, the gender difference in prevalence of asthma is statistically insignificant at the 0.05 level.

Rao and Scott (1984) offered another test procedure for contingency table analysis of complex surveys. This procedure adjusts a different chi-square test statistic and again uses an F statistic with noninteger degrees of freedom to determine the appropriate p-value. Some software packages implemented the Rao-Scott corrected statistic as the default procedure. In most situations, the Wald statistic and the Rao-Scott statistic lead to the same conclusion.

Example 15.8

Table 15.9 presents analysis of a 2 by 3 contingency table using data from Phase II of NHANES III. In this analysis, the association between vitamin use and years of education is examined with education coded into three categories (1 = less than 12 years of education; 2 = 12 years; 3 = more than 12 years). The weighted percent of vitamin users by the level of education varies from 33 percent in the first level of education to 52 percent in the third level of education. The confidence intervals for these percentages are also shown. Both the Wald and the Rao-Scott statistics are shown in this table and we draw the same conclusion from both.

We next examined the relation between the use of vitamins and the level of education for the Hispanic population. Here we used a subpopulation analysis based on the entire sample. The results are shown in Table 15.10. The estimated overall proportion of vitamin users among Hispanics is 31 percent, considerably lower than the overall value of 43 percent shown in Table 15.8. The Wald test statistic in Table 15.10 also shows there is a statistically significant relation between education and use of vitamins among Hispanics.

See **Program Note 15.4** for this analysis.

Table 15.9 Percent of vitamin use by levels of education among U.S. adults, Phase II, NHANES III (n = 9920).

	Less than H.S.	H.S. Graduate	Some College	Total
Percent	33.4	39.8	51.67	43.0
Confidence Interval	[30.1, 36.9]	[36.2, 43.5]	[47.6, 55.7]	[40.4,.45.5]
	Wald Statistic:	Chi-square (2):	51.99	
		F (2, 22):	24.87	
		p-value:	<0.0001	
	Rao-Scott Statistic:	Uncorrected chi-square (2):	234.10	
		Design-based F (1.63, 37.46):	30.28	
		p-value:	<0.0001	

Table 15.10 Percent of vitamin use by levels of education for Hispanic population, Phase II, NHANES III (n = 2593).

	Less than H.S.	H.S. Graduate	Some College	Total
Percent	26.2	32.7	44.1	30.9
Confidence Interval	[22.1, 30.7]	[28.8, 36.9]	[36.9, 51.58]	[27.1, 34.9]
	Wald Statistic:	Chi-square (2):	47.16	
		F (2, 22):	22.56	
		p-value:	<0.0001	

15.5.3 Linear and Logistic Regression Analysis

In Chapter 13, we used ordinary least squares (OLS) estimation to obtain estimates of the regression coefficients or the effects in the linear model assuming simple random sampling. However, using the OLS method with data from a complex sample design

will result in biased estimates of model parameters and their variances. Thus, confidence intervals and tests of hypotheses may be misleading.

The most widely used method of estimation for complex survey data when using the general linear model is the design-weighted least squares (DWLS) method. The DWLS approach is slightly different from the weighted least squares (WLS) method for unequal variances that derives the weights from an assumed covariance structure. In the DWLS approach, the weights come from the sampling design, and the variance/covariance is estimated using one of the methods discussed in the previous section. This approach is supported by most of the software for complex survey data analysis. Several sources provide a more detailed discussion of regression analysis of complex survey data (Korn and Graubard 1999, Section 3.5; Lohr 1999, Chapter 11).

Since these methods use the PSU total rather than the individual value as the basis for the variance computation, the degrees of freedom for this design again equal d, the number of PSUs minus the number of strata. For the test of hypothesis we need to take into account the number of parameters being tested. For example, for an F test, the numerator degrees of freedom is the number of parameters being tested (q) and the denominator degrees of freedom is $d - q + 1$.

Example 15.9

We conducted a general linear model analysis of systolic blood pressure on height, weight, age, sex (male = 0), and vitamin use (user = 1) using the same NHANES III data. We did not include any interaction terms in this example, although their inclusion would undoubtedly have increased the R-square. Imputed values were not used in this analysis. The results are shown in Table 15.11. See **Program Note 15.5** on the website for the analysis.

These results can be interpreted in the same manner as in Chapter 13. The R-square is 39 percent and the F statistic for the overall ANOVA is significant. There are five degrees of freedom for the numerator in the overall F, since five independent variables are included in the model. There are 19 (= 46 − 23 − 5 + 1, based on the numbers of PSUs, strata and independent variables in the model) (Korn and Graubard 1999) degrees of freedom for the denominator in the overall F ratio. All five explanatory variables are also individually statistically significant.

Table 15.11 Multiple regression analysis of systolic blood pressure on selected variables for U.S. adults, Phase II, NHANES III. (n = 9235).

Variable	Regression Coefficient	Standard Error	t	$p > \|t\|$	Design Effect
Height	−0.4009	0.1023	−3.92	0.001	3.39
Weight	0.0917	0.0048	19.11	<0.001	1.06
Age	0.6004	0.0132	45.58	<0.001	1.67
Sex	4.0293	0.6546	6.16	<0.001	2.44
Vitamin use	−1.1961	0.4194	−2.85	0.009	1.85
Intercept	106.2809	6.7653	15.71	<0.001	3.96
Model statistics	F (5,19):	937.30			
	p-value:	<0.0001			
	R-squared:	0.393			

In Chapter 14, we presented the logistic regression model and the maximum likelihood estimation procedure. We can also modify this estimation approach to use logistic regression with complex survey data. The modified estimation approach that incorporates the sampling weights is generally known as pseudo or weighted maximum likelihood estimation (Chambless and Boyle 1985; Roberts, Rao, and Kumar 1987). The variance/covariance matrix of the estimated coefficients is calculated by one of the methods discussed in the previous section. As discussed earlier, the degrees of freedom associated with this covariance matrix are the number of PSUs minus the number of strata. Because of all these changes to the standard approach, we use the adjusted Wald test statistic instead of the likelihood-ratio statistic in determining whether or not the model parameters, excluding the constant term, are simultaneously equal to zero.

The selection and inclusion of appropriate predictor variables for a logistic regression model can be done similarly to the process for linear regression. When analyzing a large survey data set, the preliminary analysis strategy described in the earlier section is very useful in preparing for a logistic regression analysis.

Example 15.10

Based on Phase II of NHANES III, we performed a logistic regression analysis of vitamin use on two categorical explanatory variables: sex (1 = male; 0 = female) and education (less than 12 years of education; 12 years; more than 12 years). Two dummy variables are created for the education variable: edu1 = 1 if 12 years of education and 0 otherwise; edu2 = 1 if more than 12 years and 0 otherwise; the less than 12 year category is the reference category. The results are shown in Table 15.12 (see **Program Note 15.6** for this analysis).

The log-likelihood ratio is not shown because the pseudo likelihood is used and an F statistic derived from the modified Wald statistic is shown. The numerator degrees of freedom for this statistic is 3 (based on the number of independent variables) and the denominator degrees of freedom is 21 (= 46 − 23 − 3 + 1, based the numbers of PSUs, strata, and independent variables) (Korn and Graubard 1999). The small p-value suggests that the main effects model is a significant improvement over the null model. The estimated design effects suggest that the variances of the beta coefficients are roughly twice as large as those calculated under the assumption of simple random sampling. Despite the increased standard errors, the beta coefficients for gender and education levels are significant.

Table 15.12 Logistic regression analysis of vitamin use on sex and levels of education among U.S. adults, Phase II, NHANES III (n = 9920).

Variable	Estimated Coefficient	Standard Error	t	$p > \|t\|$	Design Effect	Odds Ratio	Confidence Interval
Male	−0.4998	0.0584	−8.56	<0.001	1.96	0.61	[0.54, 0.68]
Edu1	0.2497	0.0864	2.89	0.008	2.45	1.28	[1.07, 1.53]
Edu2	0.7724	0.0888	8.69	<0.001	2.84	2.16	[1.80, 2.60]
Constant	−0.4527	0.0773	−5.86	<0.001	2.82		
Model statistics		F (3, 21):	63.61				
		p-value	<0.0001				

The rest of the results can be interpreted in the same way as in Chapter 14. The estimated odds ratio for males is 0.61, meaning that, after adjusting for education, the odds of taking vitamins for a male is 61 percent of the odds that a female uses vitamins. The 95 percent confidence interval provides a test of whether or not the odds ratio is equal to one. The odds ratio for the third level of education suggests that persons with some college education are twice as likely to take vitamins than those with less than 12 years of education for the same gender. None of the 95% confidence intervals include one, suggesting that all the effects are significant at the 0.05 level. As in regular logistic regression analysis, we may combine the estimated beta coefficients to make specific statements. For example, the estimated odds ratio for males with some college education compared with females with less than 12 years of education can be obtained by $\exp(-0.4998 + 0.7724) = 1.31$. Since we have not included any interaction effects in the model, the resulting odds ratio of 1.31 can be interpreted as indicating that the odds of taking vitamins for males with some college education is 31 percent higher than the odds for females with less than 12 years of education.

Conclusion

In this chapter, we discussed issues associated with complex sample surveys focusing on design-based statistical inference. We summarized the two key complications that arise in the analysis of data from complex sample surveys: the need to include sample weights and the need to take the sample design into account in calculating the sampling variance of weighted statistics. We presented several different approaches to the calculation of the sample variances. Practically all statistical methods discussed in previous chapters can be applied to complex survey data with some modifications. For the analysis of a specific subgroup, we pointed out that the entire sample is used although we set the weights to zero for the observations outside the subgroup. Statistical programs for complex surveys are now readily available, but one needs to guard against misuse of the programs. For a proper analysis, one must understand the sample design and conduct a thorough preliminary examination of data. We conclude this chapter by again emphasizing the need to reduce nonresponse and to study some of the nonrespondents if possible.

EXERCISES

15.1 The following data represent a small subset of a large telephone survey. The sample design was intended to be an equal probability sample on each phone number. Within each selected household one adult was sampled using the Kish selection table (Kish 1949). Some households may have more than one phone number and these households are more likely to be selected in random digit dialing. Therefore, selection probability is unequal for individual respondents.

Household	Number of Adults	Number of Phones	Smoking Status	Household	Number of Adults	Number of Phones	Smoking Status
1	3	1	yes	11	4	2	no
2	2	1	no	12	1	1	no
3	4	1	no	13	2	1	no
4	2	1	no	14	3	1	yes
5	2	1	no	15	1	1	no
6	5	2	no	16	3	1	no
7	4	1	yes	17	2	1	no
8	2	1	no	18	2	1	yes
9	3	1	yes	19	3	1	no
10	2	1	no	20	2	1	yes

Develop the sample weight for each respondent, calculate the weighted percentage of smokers, and compare with the unweighted percentage. How would you interpret the weighted and unweighted percentages?

15.2 A community mental health survey was conducted using 10 replicated samples selected by systematic sampling from a geographically ordered list of residential electric hookups (Lee et al. 1986). The total sample size was 3058, and each replicate contained about 300 respondents. The replicated samples were selected to facilitate the scheduling and interim analysis of data during a long period of screening and interviewing, not for estimating the standard errors. Because one adult was randomly selected from each household, the number of adults in each household became the sample weight for each observation. This weight was then adjusted for nonresponse and poststratification and the adjusted weights were used in the analysis. The prevalences of any mental disorders during the past six months and the odds ratios for sex differences in the six-month prevalence rates of mental disorders are shown here for the full sample and the 10 replicates.

Replicate	Prevalence Rate	Odds Ratio
Full sample	17.17	0.990
1	12.81	0.826
2	17.37	0.844
3	17.87	1.057
4	17.64	0.638
5	16.65	0.728
6	18.17	1.027
7	14.69	1.598
8	17.93	1.300
9	17.86	0.923
10	18.91	1.111

Estimate the standard errors for the prevalence rate and the odds ratio based on replicate estimates. Is the approximate standard error based on the range in replicate estimates satisfactory?

15.3 From Phase II of NHANES III, the percent of adults taking vitamin or mineral supplements was estimated to be 43.0 percent with a standard error of 1.22 percent. The design effect of this estimate was 5.98 and the sample size was 9920. What size sample would be required to estimate the same quantity with a standard error of 2 percent using a simple random sampling design?

15.4 Read the article by Gold et al. (1995). Describe how their sample was designed and selected. What was the nonresponse rate? Describe also the method of analysis. Did they account for the sampling design in their analysis? If you find any problems, how would you rectify the problems?

15.5 Using the data file extracted from the adult sample in the Phase II of NHANES III (available on the web), explore one of the following research questions and prepare a brief report describing and interpreting your analysis:

a. Are more educated adults taller than less educated people?
b. Does the prevalence rate of asthma vary by region?
c. Does the use of antacids vary by smoking status (current, previous, and never smoked)?

15.6 Read the article by Flegal et al. (1995), and prepare a critical review of it. Is the purpose and design of the survey properly integrated in the analysis and conclusion? Is the model specified appropriately? Do you think the analysis is done properly? Would you do any part of the analysis differently?

15.7 Select another research question from Exercise 15.5. Conduct the analysis with and without incorporating the weight and design features and compare the results. How would you describe the consequence of not accounting for the weight and design features in the complex survey analysis?

REFERENCES

Chambless, L. E., and K. E. Boyle. "Maximum Likelihood Methods for Complex Sample Data: Logistic Regression and Discrete Proportional Hazards Models." *Communications in Statistics — Theory and Methods*, 14:1377–1392, 1985.

Deming, W. E. *Sampling Design in Business Research*. New York: Wiley, 1960, Chapter 6.

Flegal et al. "The Influence of Smoking Cessation on the Prevalence of Overweight in the United States." *New England Journal of Medicine* 333:115–127, 1995.

Frankel, M. R. *Inference from Survey Samples*. Ann Arbor: Institute of Social Research, University of Michigan, 1971.

Gold et al. "A National Survey of the Arrangements Managed-Care Plans Make with Physicians." *New England Journal of Medicine* 333:1689–1693, 1995.

Kalton, G. *Introduction to Survey Sampling*. Beverly Hills, CA: Sage Publications (QASS, vol. 35) 1983, p. 52.

Kish, L. "A Procedure for Objective Respondent Selection within the Household." *Journal of the American Statistical Association* 44:380–387, 1949.

Kish, L. *Survey Sampling*. New York: Wiley, 1965, p. 620.

Korn, E. L., and B. I. Graubard. *Analysis of Health Surveys*. New York: Wiley, 1999.

Koch, G. G., D. H. Freeman, and J. L. Freeman. "Strategies in the Multivariate Analysis of Data from Complex Surveys." *International Statistical Review* 43:59–78, 1975.

Lee, E. S., and R. N. Forthofer. *Analyzing Complex Survey Data*, 2nd ed. Thousand Oaks, CA: Sage Publications, 2006.

Lee, E. S., R. N. Forthofer, C. E. Holzer, and C. A. Taube. "Complex Survey Data Analysis: Estimation of Standard Errors Using Pseudo-strata." *Journal of Economic and Social Measurement* 14:135–144, 1986.

Levy, P. S., and S. Lemeshow. *Sampling of Populations: Methods and Applications*, 3rd ed. New York: Wiley, 1999, Chapter 13.

Little, R. J. A., and D. B. Rubin. *Statistical Analysis with Missing Data*, 2nd ed. New York: Wiley, 2002.

Lohr, S. L. *Sampling: Design and Analysis*. New York: Duxbury Press, 1999.

McCarthy, P. J. *Replication: An Approach to the Analysis of Data from Complex Surveys*. Vital and Health Statistics, Series 2, no. 14, Fyattsville, MD: National Center for Health Statistics, 1966.

National Center for Health Statistics. *Plan and Operation of the Third National Health and Nutrition Examination Survey, 1988–94*. Vital and Health Statistics, Series 1, no. 32, Hyattsville, MD, 1994.

Plackett, R. L., and P. J. Burman. "The Design of Optimum Multi-factorial Experiments." *Biometrika* 33:305–325, 1946.

Quenouille, M. H. "Approximate Tests of Correlation in Time Series." *Journal of the Royal Statistical Society* 11(B):68–84, 1949.

Rao, J. N. K., and A. J. Scott. "On Chi-Square Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data." *Annals of Statistics* 12:46–60, 1984.

Roberts, G., J. N. K. Rao, and S. Kumar. "Logistic Regression Analysis of Sample Survey Data." *Biometrika* 74:1–12, 1987.

Sudman, S. *Applied Sampling*. New York: Academic Press, 1976.